Probability and Statistics

Cheat Sheet · v1.0.1

Flavio Schneider

ETH Zürich - D-INFK

1 Probability

1.1 Basics

Def. 1.1: Sample Space

The sample space, denoted by $\Omega \neq \emptyset$, is the set of all possible outcomes of an experiment, it can be finite or infinite.

Def. 1.2: Event

An event A is a subset of the sample space $A \subseteq \Omega$, or an element of the powerset of the sample space $A \in 2^\Omega$.

Def. 1.3: Observable Event Set

The set of all observable events is denoted by $\mathcal{F}$, where $\mathcal{F} \subseteq 2^\Omega$.

Note

- Usually if Ω is countable $\mathcal{F} = 2^\Omega$, however sometimes many events are excluded from $\mathcal{F}$ since it's not possible for them to happen.

Def. 1.4: $\sigma-$Algebra

The set $\mathcal{F}$ is called a $\sigma-$Algebra if:

(i) $\Omega \in \mathcal{F}$

(ii) $\forall A \subseteq \Omega : A \in \mathcal{F} \Rightarrow A^C \in \mathcal{F}$

(iii) $\forall (A_n)_{n\in\mathbb{N}} : A_n \in \mathcal{F} \Rightarrow \bigcup_{n=1}^\infty A_n \in \mathcal{F}$

Def. 1.5: Probability Function

$P : \mathcal{F} \to [0,1]$ is a probability function if it satisfies the following 3 axioms:

(i) $\forall A \in \mathcal{F} : P\left[A\right] \geq 0$

(ii) $P[\Omega] = 1$

(iii) $P[\bigcup_{n=1}^\infty A_n] = \sum_{n=1}^\infty P\left[A_n\right]$

where A_n are disjunct.

Properties (derived from the 3 axioms):

- $P\left[A^C\right] = 1 - P\left[A\right]$
- $P\left[\emptyset\right] = 0$
- $A \subseteq B \Rightarrow P[A] \leq P[B]$
- $P[A \cup B] = P[A] + P[B] - P[A \cap B]$

Theorem 1: Inclusion-Exclusion

Let $A_1, \dots, A_n$ be a set of events, then:

$$P\left[\bigcup_{i=1}^n A_i\right] = \sum_{k=1}^n (-1)^{k-1} S_k$$

where

$$S_k = \sum_{\substack{I \subseteq \{1,\dots,n\} \\ |I|=k}} P\left[\bigcap_{i\in I} A_i\right]$$

1.2 Discrete Probability

We talk about discrete probability if Ω is countable (finite or infinite).

Def. 1.6: Laplace Space

If $\Omega = \{\omega_1, \dots, \omega_N\}$ with $|\Omega| = N$ where all ω_i have the same probability $p_i = \frac{1}{N}$, Ω is called Laplace Space and P has a discrete unifrom distribution. For some event A we have:

$$P[A] = \frac{|A|}{|\Omega|}$$

Note

- The discrete uniform distribution exists only if Ω is finite.

1.3 Conditional Probability

Def. 1.7: Conditional Probability

Given two events A and B with $P[A] > 0$, the probability of B given A is defined as:

$$P\left[B|A\right] := \frac{P\left[B \cap A\right]}{P[A]}$$

Theorem 2: Total Probability

Let $A_1, \dots, A_n$ be a set of disjunct events $\forall i \neq j : A_i \cap A_j = \emptyset$ where $\bigcup_{i=1}^n A_i = \Omega$, then for any event $B \subseteq \Omega$:

$$P\left[B\right] = \sum_{i=1}^n P\left[B|A_i\right] P\left[A_i\right]$$

Theorem 3: Bayes' Rule

Let $A_1, \dots, A_n$ be a set of disjunct events $\forall i \neq j : A_i \cap A_j = \emptyset$ where $\bigcup_{i=1}^n A_i = \Omega$, with $P[A_i] > 0$ for all $i = 1, \dots, n$, then for an event $B \subseteq \Omega$ with $P[B] > 0$ we have:

$$P\left[A_k|B\right] = \frac{P[B|A_k]P[A_k]}{\sum_{i=1}^n P\left[B|A_i\right] P\left[A_i\right]}$$

Note

- If we have only two events A and B it simplifies to: $P[A|B] = \frac{P[B|A]P[A]}{P[B]}$

1.4 Independence

Def. 1.8: Independence

A set of events $A_1, \dots, A_n$ are independent if for all $m \in \mathbb{N}$ with $\{k_1, \dots, k_m\} \subseteq 1, \dots, n$ we have:

$$P\left[\bigcap_{i=1}^m A_{k_i}\right] = \prod_{i=1}^m P[A_{k_i}]$$

Properties

With only two events:

- A and B are independent iff $P[A \cap B] = P[A]P[B]$
- A and B are independent iff $P[B|A] = P[B]$

2 Combinatorics

Let n be the number of total objects and k be the number of object that we want to select ($k = n$ if we consider all objects), then:

Def. 2.1: Permutation

A *permutation* $P_n(k)$ is an arrangement of elements where we care about ordering.

(i) Repetition not allowed:

$$P_n(k) = \frac{n!}{(n-k)!}$$

(ii) Repetition allowed:

$$P_n(k) = n^k$$

Def. 2.2: Combination

A *combination* $C_n(k)$ is an arrangement of elements where we do *not* care about ordering.

(i) Repetition not allowed:

$$C_n(k) = \binom{n}{k} = \frac{P_n(k)}{k!} = \frac{n!}{k!(n-k)!}$$

(ii) Repetition allowed:

$$C_n(k) = \binom{n+k-1}{k}$$

Note

- Repetition is the same as replacement, since by replacing an object in the set we'll be able to use it again.

Properties

- $0! = 1$
- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- $\binom{n}{0} = \binom{n}{n} = 1$
- $\binom{n}{1} = \binom{n}{n-1} = n$
- $\binom{n}{k} = \binom{n}{n-k}$
- $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$
- $\sum_{k=0}^n \binom{n}{k} = 2^n$

3 Random Variables

3.1 Basics

Def. 3.1: Random Variable

Let $(\Omega, \mathcal{F}, P)$ be a probability space, then a *random variable* (RV) on Ω is a function:

$$X : \Omega \to \mathcal{W}(X) \subseteq \mathbb{R}$$

if the image $\mathcal{W}(X)$ is countable X is called a *discrete* random variable, otherwise it's called a *continuous* random variable.

Def. 3.2: Probability Density

The *probability density function* (PDF) $f_X : \mathbb{R} \to \mathbb{R}$ of a RV X, is function defined as:

$$f_X(x) \coloneqq P[X = x] \coloneqq P[\{\omega \mid X(\omega) = x\}]$$

with X discrete we use $p_X(t)$ instead of $f_X(t)$.

Properties

- $f_X = 0$ and $f_X \geq 0$ outside of $W(X)$.
- $\int_{-\infty}^{\infty} f_X(t)dt = 1$

Def. 3.3: Cumulative Distribution

The *cumulative distribution function* (CDF) $F_X : \mathbb{R} \to [0, 1]$ of a RV X, is a function defined as:

$$F_X(x) \coloneqq P[X \leq x] \coloneqq P[\{\omega \mid X(\omega) \leq x\}]$$

if the PDF is given it can be expressed with:

$$F_X(x) = \begin{cases} \displaystyle\sum_{x_i \leq x} p_X(x_i) & X \text{ discr.} \\ \displaystyle\int_{-\infty}^{x} f_X(t)dt & X \text{ cont.} \end{cases}$$

Properties

- *Monotone:* If $t \leq s$ then $F_X(t) \leq F_X(s)$.
- *R-continuous:* If $t > s$ then $\lim_{t \to s} F_X(t) = F_X(s)$.
- *Limits:* $\lim_{t \to -\infty} F_X(t) = 0 \wedge \lim_{t \to \infty} F_X(t) = 1$.
- $P[a < X \leq b] = F_X(b) - F_X(a) = \int_a^b f_X(t)dt$
- $P[X > t] = 1 - P[X \leq t] = 1 - F_X(t)$
- $\frac{d}{dx} F_X(x) = f_X(x)$

3.2 Expected Value

Def. 3.4: Expected Value

Let X be a RV, then the *expected value* is defined as:

$$\mathbb{E}[X] = \mu \coloneqq \begin{cases} \displaystyle\sum_{x_k \in \mathcal{W}(X)} x_k \cdot p_X(x_k) & X \text{ discr.} \\ \displaystyle\int_{-\infty}^{\infty} x \cdot f_X(x)dx & X \text{ cont.} \end{cases}$$

Properties

- $\mathbb{E}[X] \leq \mathbb{E}[Y]$ if $\forall \omega : X(\omega) \leq Y(\omega)$
- $\mathbb{E}\left[\sum_{i=0}^{n} a_i X_i\right] = \sum_{i=0}^{n} a_i \mathbb{E}[X_i]$
- $\mathbb{E}[X] = \sum_{j=1}^{\infty} P[X \geq j]$, if $W(X) \subseteq \mathbb{N}_0$.
- $\mathbb{E}\left[\sum_{i=0}^{\infty} X_i\right] \neq \sum_{i=0}^{\infty} \mathbb{E}[X_i]$
- $\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X]$
- $\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\,\mathbb{E}[Y^2]$
- $\mathbb{E}\left[\prod_{i=0}^{n} X_i\right] = \prod_{i=0}^{n} \mathbb{E}[x_i]$ for indep. $X_1, \ldots, X_n$.

Theorem 4: $\mathbb{E}$ of Functions

Let X be a RV and $Y = g(X)$, with $g : \mathbb{R} \to \mathbb{R}$, then:

$$\mathbb{E}[Y] = \begin{cases} \displaystyle\sum_{x_k \in \mathcal{W}(X)} g(x_k) \cdot p_X(x_k) & X \text{ discr.} \\ \displaystyle\int_{-\infty}^{\infty} g(x) \cdot f_X(x)dx & X \text{ cont.} \end{cases}$$

Def. 3.5: Moment-Generating Function

Let X be a RV, then the *moment-generating function* of X is defined as:

$$M_X(t) \coloneqq \mathbb{E}\left[e^{tX}\right]$$

3.3 Variance

Def. 3.6: Variance

Let X be a RV with $\mathbb{E}\left[X^2\right] < \infty$, then the *variance* of X is defined as:

$$\mathrm{Var}[X] \coloneqq \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

with the extended form:

$$\mathrm{Var}[X] = \begin{cases} \left(\displaystyle\sum_k p_X(x_k) \cdot x_k^2\right) - \mu^2 & X \text{ discr.} \\ \displaystyle\int_{-\infty}^{\infty} x^2 \cdot f_X(x)dx - \mu^2 & X \text{ cont.} \end{cases}$$

Properties

- $0 \leq \mathrm{Var}[X] \leq \mathbb{E}\left[X^2\right]$
- $\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$
- $\mathrm{Var}[aX + b] = a^2 \mathrm{Var}[X]$
- $\mathrm{Var}[X] = \mathrm{Cov}(X, X)$
- $\mathrm{Var}\left[\sum_{i=0}^{n} a_i X_i\right] =$
  $\sum_{i=0}^{n} a_i^2 \mathrm{Var}[X_i] + 2 \sum_{1 \leq i < j \leq n} a_i a_j \mathrm{Cov}(X_i, X_j)$
- $\mathrm{Var}\left[\sum_{i=0}^{n} a_i X_i\right] = \sum_{i=0}^{n} \mathrm{Var}[X_i]$
  if $\forall (i \neq j) : \mathrm{Cov}(X_i, X_j) = 0$.

Def. 3.7: Standard Deviation

Let X be a RV with $\mathbb{E}\left[X^2\right] < \infty$, then the *standard deviation* of X is defined as:

$$\sigma(X) = sd(X) \coloneqq \sqrt{\mathrm{Var}[X]}$$

3.4 Other Functions

Def. 3.8: Covariance

Let X, Y be RVs with finite expected value, then the *covariance* of X and Y is defined as:

$$\begin{aligned} \mathrm{Cov}(X, Y) &\coloneqq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \end{aligned}$$

Note

- The covariance is a measure of correlation between two random variables, $\mathrm{Cov}(X, Y) > 0$ if Y tends to increase as X increases and $\mathrm{Cov}(X, Y) < 0$ if Y tends to decrease as X increases. If $\mathrm{Cov}(X, Y) = 0$ then X and Y are uncorrelated.

Properties

- $\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X, Y)$
- $\mathrm{Cov}(X + a, Y + b) = \mathrm{Cov}(X, Y)$
- $\mathrm{Cov}(a_1 X_1 + a_2 X_2, b_1 Y_1 + b_2 Y_2) =$
  $a_1 b_1 \mathrm{Cov}(X_1, Y_1) + a_1 b_2 \mathrm{Cov}(X_1, Y_2) +$
  $a_2 b_1 \mathrm{Cov}(X_2, Y_1) + a_2 b_2 \mathrm{Cov}(X_2, Y_2)$

Def. 3.9: Correlation

Let X, Y be RVs with finite expected value, then the *correlation* of X and Y is defined as:

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}[X] \cdot \mathrm{Var}[Y]}}$$

Note

- Correlation is the same as covariance but normalized with values between -1 and 1.
- X, Y indep. $\Rightarrow \mathrm{Corr}(X, Y) = \mathrm{Cov}(X, Y) = 0$.

Def. 3.10: Indicator Function

The *indicator function* I_A for a set (event) A is defined as:

$$I_A(\omega) \coloneqq \begin{cases} 1 & w \in A \\ 0 & w \in A^C \end{cases}$$

3.5 Joint Probability

Def. 3.11: Joint PDF

The *joint probability density function* $f_{\mathbf{X}} : \mathbb{R}^n \to [0,1]$ with $\mathbf{X} = (X_1, \dots, X_n)$ is a function defined as:

$$f_{\mathbf{X}}(x_1, \dots, x_n) := P[X_1 = x_1, \dots, X_n = x_n]$$

with $\mathbf{X}$ discrete we use $p_{\mathbf{X}}(\mathbf{x})$ instead of $f_{\mathbf{X}}(\mathbf{x})$.

Def. 3.12: Joint CDF

The *joint cumulative distribution function* $F_{\mathbf{X}} : \mathbb{R}^n \to [0,1]$ with $\mathbf{X} = (X_1, \dots, X_n)$ is a function defined as:

$$F_{\mathbf{X}}(x_1, \dots, x_n) := P[X_1 \leq x_1, \dots, X_n \leq x_n]$$

if the joint PDF is given it can be expressed with:

$$F_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \sum_{t_1 \leq x_1} \cdots \sum_{t_n \leq x_n} p_{\mathbf{X}}(\mathbf{t}) & \textit{discr.} \\ \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{\mathbf{X}}(\mathbf{t}) d\mathbf{t} & \textit{cont.} \end{cases}$$

where $\mathbf{t} = (t_1, \dots, t_n)$ and $\mathbf{x} = (x_1, \dots, x_n)$.

Properties

- $\dfrac{\partial^n F_{\mathbf{X}}(x_1, \dots, x_n)}{\partial x_1, \dots, \partial x_n} = f_{\mathbf{X}}(x_1, \dots, x_n)$

Def. 3.13: Marginal PDF

The *marginal probability density function* $f_{X_i} : \mathbb{R} \to [0,1]$ of X_i given a joint PDF $f_{\mathbf{X}}(x_1, \dots, x_n)$, is defined as:

$$f_{X_i}(t_i) = \begin{cases} \sum_{t_1} \cdots \sum_{t_{i-1}} \sum_{t_{i+1}} \cdots \sum_{t_n} p_{\mathbf{X}}(\mathbf{t}) & \textit{discr.} \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{t}) d\tilde{\mathbf{t}} & \textit{cont.} \end{cases}$$

where $\tilde{\mathbf{t}} = (t_1, \dots, t_{i-1}, t_{i+1}, \dots, t_n)$, and in the discrete case $t_k \in \mathcal{W}(X_k)$.

Note

- The idea of the marginal probability is to ignore all other random variables and consider only the one we're interested to.

Def. 3.14: Marginal CDF

The *marginal cumulative distribution function* $F_{X_i} : \mathbb{R} \to [0,1]$ of X_i given a joint CDF $F_{\mathbf{X}}(x_1, \dots, x_n)$, is defined as:

$$F_{X_i}(x_i) = \lim_{x_{j \neq i} \to \infty} F_{\mathbf{X}}(x_1, \dots, x_n)$$

Def. 3.15: Conditional Distribution

The *conditional distribution* $f_{X|Y} : \mathbb{R} \to [0,1]$ is defined as:

$$\begin{aligned} f_{X|Y}(x|y) &:= P[X = x | Y = y] \\ &= \frac{P[X = x, Y = y]}{P[Y = y]} \\ &= \frac{\textit{Joint PDF}}{\textit{Marginal PDF}} \end{aligned}$$

with X and Y discrete we write $p_{X|Y}(x|y)$ instead of $f_{X|Y}(x|y)$.

3.6 Independence

Def. 3.16: Independence

The RVs $X_1, \dots, X_n$ are independent if:

$$F_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$$

similarly if their PDF is absolutely continuous they are independent if:

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

Theorem 5: Function Independence

If the RVs $X_1, \dots, X_n$ are independent where $f_i : \mathbb{R} \to \mathbb{R}$ is a function with $Y_i := f_i(X_i)$ then also $Y_1, \dots, Y_n$ are independent.

Theorem 6

The RVs $X_1, \dots, X_n$ are independent iff $\forall B_i \subseteq W(X_i)$:

$$P[X_1 \in B_1, \dots, X_n \in B_n] = \prod_{i=1}^{n} P[X_i \in B_i]$$

3.7 Joint Functions

Def. 3.17: Joint Expected Value

The *joint expected value* of a RV $Y = g(X_1, \dots, X_n) = g(\mathbf{X})$ is defined as:

$$\mathbb{E}[Y] = \begin{cases} \sum_{t_1} \cdots \sum_{t_n} g(\mathbf{t}) p_{\mathbf{X}}(\mathbf{t}) & \textit{discr.} \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{t}) f_{\mathbf{X}}(\mathbf{t}) d\mathbf{t} & \textit{cont.} \end{cases}$$

where $\mathbf{t} = (t_1, \dots, t_n)$, and in the discrete case $t_k \in \mathcal{W}(X_k)$.

Def. 3.18: Conditional Expected Value

The *conditional expected value* of RVs X, Y is:

$$\mathbb{E}[X|Y](y) = \begin{cases} \sum_{x \in \mathbb{R}} x \cdot p_{X|Y}(x|y) & \textit{discr.} \\ \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y) dx & \textit{cont.} \end{cases}$$

Properties

- $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$
- $\mathbb{E}[X|Y](y) = \mathbb{E}[X]$ if X, Y indep.

Def. 3.19

Let $Y = g(X_1, \dots, X_n) = g(\mathbf{X})$, then:

$$P[Y \in C] = \int_{A_C} f_{\mathbf{X}}(\mathbf{t}) d\mathbf{t}$$

where $A_C = \{\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n \mid g(\mathbf{x}) \in C\}$ and $\mathbf{t} = (t_1, \dots, t_n)$.

Theorem 7: Transformation

Let F be continuous and a strictly increasing CDF and let $X \sim \mathcal{U}(0,1)$, then:

$$Y = F^{-1}(X) \Rightarrow F_Y = F$$

3.8 Evaluation

Guide 3.1: Monte Carlo Integration

Let $I = \int_a^b g(x)dx$ be the integral of a function that is hard to evaluate, then:

$$\begin{aligned} I &= \int_a^b g(x)dx \\ &= (b-a) \int_a^b g(x) \frac{1}{b-a} dx \\ &= (b-a) \int_{-\infty}^{\infty} g(x) f_{\mathcal{U}}(x) dx \\ &= (b-a) \cdot \mathbb{E}[g(\mathcal{U})] \end{aligned}$$

where $\mathcal{U}(a,b)$ is uniformely distributed. Then by the LLN know that we can approximate $\mathbb{E}[g(\mathcal{U})]$ by randomly sampling $u_1, u_2, \dots$ from $\mathcal{U}(a,b)$.

$$\frac{b-a}{n} \sum_{i=1}^{n} g(u_i) \xrightarrow[n \to \infty]{} (b-a) \cdot \mathbb{E}[g(\mathcal{U})]$$

Guide 3.2: Transformation

If we have a RV X with known CDF (strictly increasing) with $Y = g(X)$, to evaluate F_Y and f_Y we proceed as follows:

(i) $F_Y(t) = P[g(X) \leq t] = \int_{A_g} f_X(s) ds$

(ii) $f_Y(t) = \frac{dF_Y(t)}{dt}$

where $A_g = \{s \in \mathbb{R} \mid g(s) \leq t\}$

Guide 3.3: Sum Convolution

Let $X_1, \dots, X_n$ be independent RVs then the sum $Z = X_1 + \cdots + X_n$ has a PDF $f_Z(z)$ evaluated with a convolution between all PDFs:

$$f_Z(z) = (f_{X_1}(x_1) * \cdots * f_{X_n}(x_n))(z)$$

in the special case that $Z = X + Y$:

$$f_Z(z) = \begin{cases} \sum_{x_k \in \mathcal{W}(X)} p_X(x_k) p_Y(z - x_k) & \textit{discr.} \\ \int_{-\infty}^{\infty} f_X(t) f_Y(z-t) dt & \textit{cont.} \end{cases}$$

Note

- Often is much easier to use properties of the RVs to find the sum instead of evaluating the convolution.

Guide 3.4: Product

Let X, Y be independent RVs then to evaluate the PDF and CDF of $Z = XY$ we proceed as follows:

$$\begin{aligned} F_Z(z) &= P[XY \leq z] \\ &= P\left[X \geq \tfrac{z}{Y}, Y < 0\right] + P\left[X \leq \tfrac{z}{Y}, Y > 0\right] \\ &= \int_{-\infty}^{0} \left[\int_{\frac{z}{y}}^{\infty} f_X(x)dx\right] f_Y(y)dy \\ &+ \int_{0}^{\infty} \left[\int_{-\infty}^{\frac{z}{y}} f_X(x)dx\right] f_Y(y)dy \end{aligned}$$

where the PDF is:

$$f_Z(z) = \frac{d}{dz}F_Z(z) = \int_{-\infty}^{\infty} f_Y(y) f_X\left(\frac{z}{y}\right)\frac{1}{|y|}dy$$

Guide 3.5: Quotient

Let X, Y be independent RVs then to evaluate the PDF and CDF of $Z = \frac{X}{Y}$ we proceed as follows:

$$\begin{aligned} F_Z(z) &= P\left[\frac{X}{Y} \leq z\right] \\ &= P[X \geq zY, Y < 0] + P[X \leq zY, Y > 0] \\ &= \int_{-\infty}^{0} \left[\int_{yz}^{\infty} f_X(x)dx\right] f_Y(y)dy \\ &+ \int_{0}^{\infty} \left[\int_{-\infty}^{yz} f_X(x)dx\right] f_Y(y)dy \end{aligned}$$

where the PDF is:

$$f_Z(z) = \frac{d}{dz}F_Z(z) = \int_{-\infty}^{\infty} |y|\, f_X(yz) f_Y(y)dy$$

3.9 Sum and Average

Let $X_1, \ldots, X_n$ be i.i.d RVs with finite mean μ, standard deviation σ, and let Z_n be the *standardization* of a RV Y defined as:

	Sum	*Average*
Y	$S_n = \sum_{i=1}^{n} X_i$	$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
$\mathbb{E}[Y]$	$n\mu$	μ
$\text{Var}[Y]$	$n\sigma^2$	$\frac{\sigma^2}{n}$
$\sigma(Y)$	$\sqrt{n}\sigma$	$\frac{\sigma}{\sqrt{n}}$
Z_n	$\frac{S_n - n\mu}{\sigma\sqrt{n}}$	$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$

3.10 Convergence

Def. 3.20: Probability Convergence

Let $X_1, X_2, \ldots$ and Y be RV on the same probability space, then:

(i) $X_1, X_2, \ldots$ converges to Y in prob. if:

$$\forall \epsilon > 0 \lim_{n\to\infty} P[|X - Y| > \epsilon] = 0$$

(ii) $X_1, X_2, \ldots$ converges to Y in L^p for $p > 0$ if:

$$\lim_{n\to\infty} \mathbb{E}[|X_n - Y|^p] = 0$$

(iii) $X_1, X_2, \ldots$ converges to Y, P-almost surely if:

$$P\left[\lim_{n\to\infty} X_n = Y\right] = P\left[\left\{w \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = Y(\omega)\right\}\right] = 1$$

Def. 3.21: Distribution Convergence

Let $X_1, X_2, \ldots$ and Y be RV, with CDF $F_{X_1}, F_{X_2}, \ldots$ and F_Y then $X_1, X_2, \ldots$ converges to Y in distribution if:

$$\forall x \in \mathbb{R} \lim_{n\to\infty} F_{X_n}(x) = F_Y(x)$$

3.11 Inequalities

Theorem 8: Markov-Inequality

Let X be a RV and $g : \mathcal{W}(X) \to [0, \infty)$ be an increasing function, then for all c with $g(c) > 0$ we have:

$$P[X \geq c] \leq \frac{\mathbb{E}[g(X)]}{g(c)}$$

Note: for practical uses usually $g(x) = x$.

Theorem 9: Chebyshev-Inequality

Let X a RV with $\text{Var}[X] < \infty$ then if $b > 0$:

$$P[|X - \mathbb{E}[X]| \geq b] \leq \frac{\text{Var}[X]}{b^2}$$

Theorem 10

Let $X_1, \ldots, X_n$ *i.i.d.* where $\forall t : M_X(t) < \infty$ then for any $b \in \mathbb{R}$:

$$P[S_n \geq b] \leq \exp\left(\inf_{t\in\mathbb{R}}(n \log M_X(t) - tb)\right)$$

Theorem 11: Chernoff-Inequality

Let $X_1, \ldots, X_n$, with X_i *i.i.d* $\sim \text{Be}(p_i)$ and $S_n = \sum_{i=1}^{n}$ where $\mu_n := \mathbb{E}[S_n] = \sum_{i=1}^{n} p_i$ then if $\delta > 0$:

$$P[S_n \geq (1+\delta)\mu_n] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\mu_n} \approx \mathcal{O}(e^{-n})$$

3.12 Limit Theorems

Theorem 12: Law of Large Numbers

Let $X_1, X_2, \ldots$ be i.i.d RVs with finite mean μ. Let $\overline{X}_n$ be the average of the first n variables, then the *law of large numbers* (LLN) says that (different versions):

(i) *Weak*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} \xrightarrow[n\to\infty]{} \mu$$

(ii) *Weak*

$$\forall \epsilon\ P\left[|\overline{X}_n - \mu| > \epsilon\right] \xrightarrow[n\to\infty]{} 0$$

(iii) *Weak*

$$\forall \epsilon\ P\left[|\overline{X}_n - \mu| < \epsilon\right] \xrightarrow[n\to\infty]{} 1$$

(iv) *Strong*

$$P\left[\left\{\omega \in \Omega \mid \overline{X}_n(\omega) \xrightarrow[n\to\infty]{} \mu\right\}\right]$$

Note

- The law of large numbers says that if we average n i.i.d. RV, then the more n increases the more the average is probable to be close to the expected value of the RVs: $\overline{X}_n \approx \mu$.

Properties

- $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(X_i) = \mathbb{E}[f(X)]$

Theorem 13: Central Limit Theorem

Let $X_1, \ldots, X_n$ be i.i.d RVs with finite mean μ and standard deviation σ. Let Z_n be a standardization, then for any $z \in \mathbb{R}$:

$$\lim_{n\to\infty} F_{Z_n}(z) = \lim_{n\to\infty} P[Z_n \leq z] = \Phi(z)$$

Where a practical application is that for n big:

(i) $P[Z_n \leq z] \approx \Phi(z)$
(ii) $Z_n \approx \mathcal{N}(0, 1)$
(iii) $S_n \approx \mathcal{N}(n\mu, n\sigma^2)$
(iv) $\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

Note

- The idea is that any (normalized) sum or average of RVs approaches a (standard) normal distribution as n gets bigger.

4 Estimators

4.1 Basics

Let $X_1, \dots, X_n$ i.i.d. RVs, drawn according to some distribution P_θ parametrized by $\theta = (\theta_1, \dots, \theta_m) \in \Theta$ where Θ is the set of all possible parameters for the selected distribution. Then the goal is to find the best estimator $\hat{\theta} \in \Theta$ such that $\hat{\theta} \approx \theta$ since the real θ cannot be known exactly from a finite sample.

Def. 4.1: Estimator

An *estimator* $\hat{\theta}_j$ for a parameter θ_j is a RV $\hat{\theta}_j(X_1, \dots, X_n)$ that is symbolized as a function of the observed data.

Def. 4.2: Estimate

An *estimate* $\hat{\theta}_j(x_1, \dots, x_n)$ is a realization of the estimator RV, it's real value for the estimated parater.

Def. 4.3: Bias

The *bias* of an estimator $\hat{\theta}$ is defined as:

$$\mathrm{Bias}_\theta[\hat{\theta}] := \mathbb{E}_\theta[\hat{\theta}] - \theta = \mathbb{E}_\theta[\hat{\theta} - \theta]$$

we say that an estimator is *unbiased* if:

$$\mathrm{Bias}_\theta[\hat{\theta}] = 0 \quad \text{or} \quad \mathbb{E}_\theta[\hat{\theta}] = \theta$$

Def. 4.4: Mean Squared Error

The *mean squared error* (MSE) of an estimator $\hat{\theta}$ is defined as:

$$\mathrm{MSE}_\theta[\hat{\theta}] := \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathrm{Var}_\theta[\hat{\theta}] + (\mathbb{E}_\theta[\hat{\theta}] - \theta)^2$$

Def. 4.5: Consistent

A squence of estimators $\hat{\theta}^{(n)}$ of the parameter θ is called *consistent* if for any $\epsilon > 0$:

$$P_\theta[|\hat{\theta}^{(n)} - \theta| > \epsilon] \xrightarrow[n \to \infty]{} 0$$

Note

- The idea is that an estimator is consistent only if as the sample data increases the estimator approaches the real parameter.

4.2 Maximum-Likelihood Method

Def. 4.6: Likelihood Function

The *likelihood function* L is defined as:

$$L(x_1, \dots, x_n; \theta) = \begin{cases} p(x_1, \dots, x_n; \theta) & \textit{discr.} \\ f(x_1, \dots, x_n; \theta) & \textit{cont.} \end{cases}$$

Def. 4.7: MLE

The *maximum likelihood estimator* $\hat{\theta}$ for θ is defined as:

$$\hat{\theta} \in \left\{ \underset{\theta \in \Theta}{\arg\max}\, L(X_1, \dots, X_n; \theta) \right\}$$

Guide 4.1: Evaluation

Given a i.i.d. sample of data $x_1, \dots, x_n$ and a distribution P_θ:

(i) Identify the parameters $\theta = (\theta_1, \dots, \theta_m)$ for the given distribution (e.g. if normal $\theta = (\theta_1 = \mu, \theta_2 = \sigma^2)$).

(ii) Find the log likelihood, we use the log of the likelihood since it's much easier to differentiate afterwards, and the maximum of L is preserved ($\forall \theta_j$):

$$\begin{aligned} g(\theta_j) &:= \log L(x_1, \dots, x_n; \theta_j) \\ &= \log \prod_{i=1}^{n} f(x_i; \theta_j) \end{aligned}$$

the goal here is to split f into as many sums as possible using log properties (easier to differentiate).

(iii) Find the maximum of the log likelihood, note that if the distribution is simple it might be easier to use the normal likelihood function and manually find the max, and if the distribution is hard we might have to use iterative methods instead of differentiation. Then for each parameter θ_j:

$$\frac{dg}{d\theta_j} \overset{\text{MAX}}{=} 0$$

Often we want to find inside the derivative set to 0 a sum or average $(S_n, \overline{X}_n)$.

(iv) State the final MLE, where each parameter estimator is the max found for θ_j:

$$\hat{\theta}_{MLE} = (\hat{\theta}_1, \dots, \hat{\theta}_m)$$

4.3 Method of Moments

Def. 4.8: Theoretical Moments

Let X be a RV, then:

(i) The k^{th} *moment* of X is: $\mu_k := m_k = \mathbb{E}[X^k]$

(ii) The k^{th} *central moment* of X is: $\mu_k^* := m_k^* = \mathbb{E}[(X - \mu)^k]$

(iii) The k^{th} *absolut moment* of X is: $M_k := \mathbb{E}[|X|^k]$ *(not used for MOM)*

Def. 4.9: Sample Moments

Let X be a RV, then given a sample $x_1, \dots, x_n$ using the Law of Large numbers:

(i) The k^{th} *moment* is evaluated as:

$$\hat{\mu}_k(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

(ii) The k^{th} *central moment* is evaluated as:

$$\hat{\mu}_k^*(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_1)^k$$

Guide 4.2: Evaluation

Given a i.i.d. sample of data $x_1, \dots, x_n$ and a distribution P_θ:

(i) Identify the parameters $\theta = (\theta_1, \dots, \theta_m)$ for the given distribution.

(ii) Since the distribution is given the expected value $\mathbb{E}_\theta[X] = g_1(\theta_1, \dots, \theta_m)$ and variance $\mathrm{Var}_\theta[X] = g_2(\theta_1, \dots, \theta_n)$ are known. The functions g_i with $0 \le i \le m$ are parametrized by θ and each of them is equal to a *thoretical* moment.

(iii) Since we have also the sample data to work with we can equate the theortical moments to the moment estimators:

$$\begin{aligned} g_1(\theta_1, \dots, \theta_m) &= \hat{\mu}_1(x_1, \dots, x_n) \\ g_2(\theta_1, \dots, \theta_m) &= \hat{\mu}_2^*(x_1, \dots, x_n) \\ &\vdots \\ g_i(\theta_1, \dots, \theta_m) &= \hat{\mu}_i^*(x_1, \dots, x_n) \\ &\vdots \\ g_m(\theta_1, \dots, \theta_m) &= \hat{\mu}_m^*(x_1, \dots, x_n) \end{aligned}$$

(iv) Now since there are m equations and m unknown thetas we can solve for each θ and set it as the estimator.

$$\hat{\theta}_{MOM} = (\hat{\theta}_1, \dots, \hat{\theta}_m)$$

Note

- The first moment is the expected value, estimated with: $\hat{\mu}_1(x_1, \dots, x_n) = \overline{x}_n$ (average) and the second *central* moment is the variance, estimated with: $\hat{\mu}_2^*(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2$. Note that we always use the *central* moments for $i > 1$.
- If we are given only the PDF of a distribution we can still evaluate the theoretical moments by solving the expected value integral (or summation if discrete).
- To check if $\hat{\theta}_i$ is unbiased we solve $\mathbb{E}_\theta[\hat{\theta}_i]$ (parametrized by θ is important) and check whether it equals θ.

Properties

Useful to simplify MLM:

- $\prod_{i=1}^{n} a \cdot x_i = a^n \prod_{i=1}^{n} x_i$
- $\log\left(\prod_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} \log(x_i)$
- $\log\left(\sum_{i=1}^{n} e^{a \cdot x_i}\right) = a \sum_{i=1}^{n} x_i$

5 Hypothesis Testing

Let $X_1, \ldots, X_n$ i.i.d. RVs, is distributed according to some distribution P_θ parametrized by $\theta = (\theta_1, \ldots, \theta_m) \in \Theta$ where $\Theta = \Theta_0 \cup \Theta_A$ is the set of all possible parameters for the selected distribution divided in two distinct subsets $\Theta_0 \cap \Theta_A = \emptyset$. Then the goal is to *test* wheter the unknown θ lies inside Θ_0 or Θ_A, this decision system is written as $H_0 : \theta \in \Theta_0$ (*null hypothesis*) and $H_A : \theta \in \Theta_A$ (*alternative hypothesis*).

Def. 5.1: Test

Concretely a *test* is composed of a function of the sample $t(x_1, \ldots, x_n) = t$ and a *rejection region* $K \subseteq \mathbb{R}$. The decision of the test is then written as RV:

$$I_{t \in K} = \begin{cases} 1, & t \in K: \ \textit{reject } H_0 \\ 0, & t \notin K: \ \textit{do not reject } H_0 \end{cases}$$

Def. 5.2: Test Statistic

The *test statistic* $T(X_1, \ldots, X_n)$ is a RV, it is distributed according to some standard statistic (z, t, χ^2).

5.1 Steps

(i) **Model**: identify the model P_θ, or which distribution does X_i *i.i.d.* $\sim P_\theta$ follow and what are the known and unknown parameters of θ.

(ii) **Hypothesis**: identify the null and alternative hypothesis, in the null hypothesis we should explicitely state the parameters value given.

(iii) **Statistic**: identify the test statistic T of H_0 and H_A based on the sample size n and the amount of known parameteres of P_θ.

(iv) **H_0 Statistic**: state the distribution of the test statistic under H_0.

(v) **Rejection Region**: based on the test statistic and the significance level α evaluate the rejection region K.

(vi) **Result**: based on the observed data and the rejection region reject H_0 or don't reject H_0.

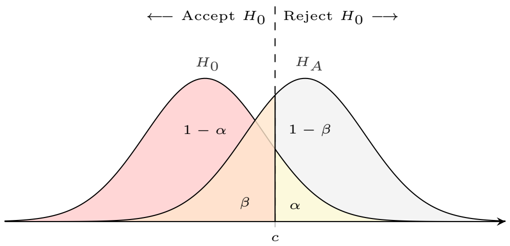
(vii) **Errors (optional)**: compute the probability of error, significance and power to decide how reliable is the test result.

5.2 Hypotheses

To test an hypothesis we must establish the null H_0 and alternative H_A hypotheses. The null hypothesis is the default set of parameters θ, or what we expect to happen if our experiment fails and the alternative hypothesis is rejected.

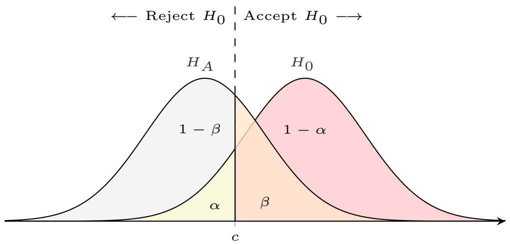
Right-Tailed (RT)

$$H_0 : \theta = \theta_0, \quad H_A = \theta > \theta_0$$



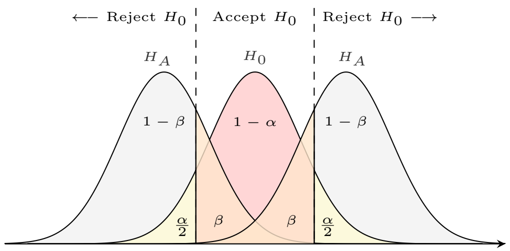
Left-Tailed (LT)

$$H_0 : \theta = \theta_0, \quad H_A = \theta < \theta_0$$



Two-Tailed (TT)

$$H_0 : \theta = \theta_0, \quad H_A = \theta \neq \theta_0$$



5.3 Statistic

X_i	n	σ^2	*Statistic*
$\mathcal{N}(\mu, \sigma^2)$	any	known	z-Test
$\mathcal{N}(\mu, \sigma^2)$	small	unknown	t-Test
any	any	any	LR-Test

LR-Test

Def. 5.3: Likelihood-Ratio

Let $L(x_1, \ldots, x_n; \theta)$ be the likelhood function where $\theta_0 \in \Theta_0$ and $\theta_A \in \Theta_A$, then the *Likelihood-Ratio* is defined as:

$$R(x_1, \ldots, x_n; \theta_0, \theta_A) := \frac{L(x_1, \ldots, x_n; \theta_0)}{L(x_1, \ldots, x_n; \theta_A)}$$

Note

- The intuition is that the likelihood function will tend to be the highest near the true value of θ, thus by evaluating the Likelihood-Ratio R between θ_0 and θ_A we can conclude that if $R < 1$ the probability of getting the observed data is higher under H_A where if $R > 1$ the probability of getting the obeserved data is higher under H_0.

Theorem 14: Neyman-Pearson

Let $T := R(x_1, \ldots, x_n; \theta_0, \theta_A)$ be the test statistic, $K := [0, c)$ be the rejection region and $\alpha^* := P_{\theta_0}[T \in K] = P_{\theta_0}[T < c]$. Then for any other test (T', K') with $P_{\theta_0}[T' \in K'] \leq \alpha^*$ we have:

$$P_{\theta_A}[T' \in K'] \leq P_{\theta_A}[T \in K]$$

Note

- The idea of the lemma is that making a decision based on the Likelihood-Ratio Test with T and K will maximise the power of the test, any other test will have a smaller power. Thus given a fixed α^*, this is *the best* way to do hypothesis testing.

z-Test

Def. 5.4: z-Test

The *z-test* is used when the data follows a normal distribution and σ^2 is known.

(i) *Statistic Under H_0*:

$$T = \frac{\overline{X_n} - \mu_0}{\sigma / \sqrt{n}} \sim \mathcal{N}(0, 1)$$

(ii) *Rejection Region*:

- $K \overset{\text{RT}}{=} [z_{1-\alpha}, \infty)$
- $K \overset{\text{LT}}{=} (-\infty, z_\alpha]$
- $K \overset{\text{TT}}{=} (-\infty, z_{\frac{\alpha}{2}}] \cup [z_{1-\frac{\alpha}{2}}, \infty)$

Properties

- $\Phi^{-1}(\alpha) = z_\alpha = -z_{1-\alpha}$
- $z_{0.95} = 1.645$, $z_{0.975} = 1.960$

t-Test

Def. 5.5: t-Test

The *t-test* is used when the data follows a normal distribution, n is small (usually $n < 30$) and σ^2 is unknown.

(i) *Statistic Under H_0*:

$$T = \frac{\overline{X_n} - \mu_0}{S / \sqrt{n}} \sim t(n-1)$$

where $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X_n})^2$

(ii) *Rejection Region*:

- $K \overset{\text{RT}}{=} [t_{n-1,1-\alpha}, \infty)$
- $K \overset{\text{LT}}{=} (-\infty, t_{n-1,\alpha}]$
- $K \overset{\text{TT}}{=} (-\infty, t_{n-1,\frac{\alpha}{2}}] \cup [t_{n-1,1-\frac{\alpha}{2}}, \infty)$

Properties

- $t_{m,\alpha} = -t_{m,1-\alpha}$

Two-Sample Tests

Def. 5.6: Paried Two-Sample Test

The *paried two-sample test* is used when we have $Y_1, \ldots, Y_n$ *i.i.d.* $\sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and $Z_1, \ldots, Z_n$ *i.i.d.* $\sim \mathcal{N}(\mu_Z, \sigma_Z^2)$ and $X_i = Y_i - Z_i$, then $X_1, \ldots, X_n$ *i.i.d.* $\sim \mathcal{N}(\mu_Y - \mu_Z, \sigma = \sigma_Y^2 - \sigma_Z^2)$, thus if σ is known we proceed with a z-test on X otherwise with a t-test on X.

Def. 5.7: Unpaired Two-Sample Test

The *unparied two-sample* test is used when we have $X_1, \ldots, X_n$ *i.i.d.* $\sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_n$ *i.i.d.* $\sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ for X_i, Y_j independent.
For known σ_X, σ_Y:

(i) *Hypothesis*: $H_0 : \mu_X - \mu_Y = \mu_0$
(ii) *Statistic Under* H_0:

$$T = \frac{\overline{X}_n - \overline{Y}_n - \mu_0}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim \mathcal{N}(0,1)$$

(iii) *Rejection Region*:

- $K \overset{\text{RT}}{=} [z_{1-\alpha}, \infty)$
- $K \overset{\text{LT}}{=} (-\infty, z_\alpha]$
- $K \overset{\text{TT}}{=} (-\infty, z_{\frac{\alpha}{2}}] \cup [z_{1-\frac{\alpha}{2}}, \infty)$

For unknown $\sigma_X = \sigma_Y > 0$:

(i) *Hypothesis*: $H_0 : \mu_X - \mu_Y = \mu_0$
(ii) *Statistic Under* H_0:

$$T = \frac{\overline{X}_n - \overline{Y}_n - \mu_0}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

(iii) *Rejection Region* ($d := n + m - 2$):

- $K \overset{\text{RT}}{=} [t_{d,1-\alpha}, \infty)$
- $K \overset{\text{LT}}{=} (-\infty, t_{d,\alpha}]$
- $K \overset{\text{TT}}{=} (-\infty, t_{d,\frac{\alpha}{2}}] \cup [t_{d,1-\frac{\alpha}{2}}, \infty)$

5.4 Errors, Significance, Power

We use the test statistic T distributed according to P_θ to evaluate the probability of errors:

H_0	Don't Reject ($T \notin K$)	Reject ($T \in K$)
true	Correct Decision	Type 1 Error (α) False Alarm False Positive
false	Type 2 Error (β) Missed Alarm False Negative	Correct Decision

Probabilities:

- **Type 1 Error**
  $P[T \in K \mid H_0 \textit{ true}] = P_{\theta_0}[T \in K] = \alpha$
- **Type 2 Error**
  $P[T \notin K \mid H_0 \textit{ false}] = P_{\theta_A}[T \notin K] = \beta$
- **Significance Level**
  $P[T \notin K \mid H_0 \textit{ true}] = P_{\theta_0}[T \notin K] = 1 - \alpha$
- **Power**
  $P[T \in K \mid H_0 \textit{ false}] = P_{\theta_A}[T \in K] = 1 - \beta$

Note:

- The significance level should be small (near 0) and the power large (near 1).
- Smaller $\alpha \Rightarrow$ Smaller power.

5.5 P-Value

Def. 5.8: P-Value

The *p-value* is the probability of getting the observed value of the test statistic $T(\omega) = t(x_1, \ldots, x_n)$, or a value with even greater evidence against H_0, if the null hypothesis is actually true.

- *p-value* $\overset{\text{RT}}{=} P_{\theta_0}[T \geq T(\omega)]$
- *p-value* $\overset{\text{LT}}{=} P_{\theta_0}[T \leq T(\omega)]$
- *p-value* $\overset{\text{TT}}{=} P_{\theta_0}[|T| \geq T(\omega)]$

Note

- We can then still decide the test and reject H_0 if *p-value* $< \alpha$ ($\alpha = 0.01$ very strong evidence, $\alpha = 0.05$ strong evidence, $\alpha > 0.1$ weak evidence).
- The *p-value* can also be viewed as the smallest α^* such that H_0 is rejected given the observed value of the test statistic $t(x_1, \ldots, x_n)$.

5.6 Confidence Interval

Def. 5.9: Confidence Interval

Given α (type-1 error) and an unknown parameter θ the *confidence interval* $C(X_1, \ldots, X_n) := [a, b]$ tells us that with probability at least $1 - \alpha$ the real parameter θ is contained in C ($\theta \in C$). Evaluated as:

$$\begin{aligned} 1 - \alpha &\leq P_\theta[\theta \in C(X_1, \ldots, X_n)] \\ &= P_\theta[a < \theta < b] \end{aligned}$$

Where a and b are:

(i) *For* $\theta := \mu$ *and known* σ*:*
$a := \overline{X}_n - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
$b := \overline{X}_n + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

(ii) *For* $\theta := \mu$ *and unknown* σ*:*
$a := \overline{X}_n - t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$
$b := \overline{X}_n + t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$

(iii) *For* $\theta := \sigma^2$ *and unknown* μ, σ*:*
$a := \frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}$
$b := \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}$

6 Discrete Distributions

6.1 Discrete Uniform Distribution

Notation	$X \sim \mathcal{U}(a,b)$
Experiment	What is the probability that we pick the value x knowing that all $n = b-a+1$ values between a and b are equally likely to be picked?
Support	$x \in \{a, a+1, \dots, b-1, b\}$
$p_X(x)$	$\frac{1}{n}$
$F_X(x)$	$\frac{x-a+1}{n}$
$\mathbb{E}[X]$	$\frac{a+b}{2}$
$\mathrm{Var}[X]$	$\frac{(b-a+1)^2-1}{12}$

6.2 Bernulli Distribution

Notation	$X \sim \mathrm{Be}(p)$
Experiment	What is the probability of success or failure is success has probability p?
Support	$x \in \{0, 1\}$
$p_X(x)$	$\begin{cases} 1-p & x=0 \\ p & x=1 \end{cases}$
$F_X(x)$	$\begin{cases} 0 & x<0 \\ 1-p & 0 \le x \le 1 \\ 1 & x>1 \end{cases}$
$\mathbb{E}[X]$	p
$\mathrm{Var}[X]$	$p(1-p)$

6.3 Binomial Distribution

Notation	$X \sim \mathrm{Bin}(n,p)$
Experiment	What is the probability of x successes in n trials if one success has probability p?
Support	$x \in \{0, 1, \dots, n\}$
$p_X(x)$	$\binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$
$F_X(x)$	$\sum_{i=1}^{x} p_X(i)$
$\mathbb{E}[X]$	np
$\mathrm{Var}[X]$	$np(1-p)$

Properties

- *Poisson Approximation*: If $X \sim \mathrm{Bin}(n,p)$ and $n \gg 0$, $np < 5$, then $X \sim \mathrm{Poi}(np)$.
- *Normal Approximation*: If $X \sim \mathrm{Bin}(n,p)$ and $n \gg 0$, $np > 5$, $n(1-p) > 5$ with $p = P[a < X \le b]$, then: $p \approx \Phi\left(\frac{b+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$.

6.4 Geometric Distribution

Notation	$X \sim \mathrm{Geo}(p)$
Experiment	What is the probability of one success in x trials if one success has probability p?
Support	$x \in \{1, 2, \dots\}$
$p_X(x)$	$(1-p)^{x-1} \cdot p$
$F_X(x)$	$1-(1-p)^x$
$\mathbb{E}[X]$	$\frac{1}{p}$
$\mathrm{Var}[X]$	$\frac{1-p}{p^2}$

Properties

- *Memoryless:* $P[X > m+n \mid X \ge m] = P[X > n]$
- *Sum:* $(\sum_{i=1}^{n} X_i \sim \mathrm{Geo}(p)) \sim \mathrm{NB}(n,p)$

6.5 Negative Binomial Distribution

Notation	$X \sim \mathrm{NB}(r,p)$
Experiment	What is the probability of r successes in x trials if one success has probability p?
Support	$x \in \{r, r+1, r+2, \dots\}$
$p_X(x)$	$\binom{x-1}{r-1} \cdot (1-p)^{x-r} \cdot p^r$
$F_X(x)$	$\sum_{i=1}^{x} p_X(i)$
$\mathbb{E}[X]$	$\frac{r}{p}$
$\mathrm{Var}[X]$	$\frac{r(1-p)}{p^2}$

6.6 Hypergeometric Distribution

Notation	$X \sim \mathrm{HGeom}(n,m,r)$
Experiment	What is the probability of picking x elements of *type 1* out of m, if there are r elements of *type 1* and $n-r$ elements of type *type 2*?
Support	$x \in \{1, 2, \dots, \min(m,r)\}$
$p_X(x)$	$\binom{r}{x}\binom{n-r}{m-x} \Big/ \binom{n}{m}$
$F_X(x)$	$\sum_{i=1}^{x} p_X(i)$
$\mathbb{E}[X]$	$\frac{rm}{n}$
$\mathrm{Var}[X]$	$\frac{(n-r)nm(n-m)}{(2n-r)^2(n-1)}$

Note

- The items are picked *without* replacement.

6.7 Poisson Distribution

Notation	$X \sim \mathrm{Poi}(\lambda)$
Experiment	What is the probability that x events happen in one unit of time knowing that on average λ events happen on one unit of time?
Support	$x \in \{0, 1, \dots\} = \mathbb{N}_0$
$p_X(x)$	$e^{-\lambda} \frac{\lambda^x}{x!}$
$F_X(x)$	$e^{-\lambda} \sum_{i=0}^{x} \frac{\lambda^i}{i!}$
$\mathbb{E}[X]$	λ
$\mathrm{Var}[X]$	λ

Properties

- Let $X = \sum_{i=1}^{n} X_i \sim \mathrm{Poi}(\lambda_i)$ where X_i are independend, then $X \sim \mathrm{Poi}\left(\sum_{i=1}^{n} \lambda_i\right)$
- If $X = c + Y$ and $Y \sim \mathrm{Poi}(\lambda)$ then $X \sim \mathrm{Poi}(\lambda)$.

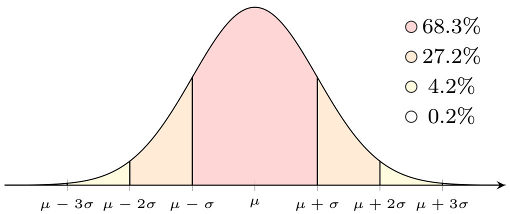
7 Continuous Distributions

7.1 Uniform Distribution

Notation	$X \sim \mathcal{U}(a,b)$
Experiment	What is the probability that we pick the value x knowing that all values between a and b are equally likely to be picked?
Support	$x \in [a,b]$
$f_X(x)$	$\begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{else} \end{cases}$
$F_X(x)$	$\begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$
$\mathbb{E}[X]$	$\dfrac{a+b}{2}$
$\mathrm{Var}[X]$	$\dfrac{(b-a)^2}{12}$

7.2 Normal Distribution

Notation	$X \sim \mathcal{N}(\mu, \sigma^2)$
Experiment	What is the probability that we pick the number x knowing that all values have a mean of μ and a standard deviation of σ?
Support	$x \in \mathbb{R}$
$f_X(x)$	$\dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
$F_X(x)$	$\Phi\left(\frac{x-\mu}{\sigma}\right)$, *(use table)*
$\mathbb{E}[X]$	μ
$\mathrm{Var}[X]$	σ^2



Properties

- $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and $Z \sim \mathcal{N}(\mu_Z, \sigma_Z^2)$ then $X + Y \sim \mathcal{N}(\mu_Y + \mu_Z, \sigma_Y^2 + \sigma_Z^2)$

7.3 Exponential Distribution

Notation	$X \sim \mathrm{Exp}(\lambda)$
Experiment	What is the probability that there are x units of time until the next event, knowing that on average λ events happen in one unit of time?
Support	$x \in [0, \infty)$
$f_X(x)$	$\begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases}$
$F_X(x)$	$\begin{cases} 1 - e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases}$
$\mathbb{E}[X]$	$\dfrac{1}{\lambda}$
$\mathrm{Var}[X]$	$\dfrac{1}{\lambda^2}$

Properties

- *Memoryless:* $P[X > m + n \mid X \ge m] = P[X > n]$

7.4 Gamma Distribution

Notation	$X \sim \mathrm{Ga}(\alpha, \lambda)$
Experiment	What is the probability that there are x units of time until the next α events, knowing that on average λ events happen in one unit of time?
Support	$x \in \mathbb{R}^+$
$f_X(x)$	$\begin{cases} \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases}$
$F_X(x)$	$\int_0^x f_X(t)dt$
$\mathbb{E}[X]$	$\dfrac{\alpha}{\lambda}$
$\mathrm{Var}[X]$	$\dfrac{\alpha}{\lambda^2}$

Note

- The gamma function $\Gamma(z)$ is the continuous analogous of the factorial: $\Gamma(n) = (n-1)!$ for $n > 0$, and is defined as $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$.

Properties

- If $X = \sum_{i=1}^{\alpha} Y_i$ with Y_i *i.i.d.* $\sim \mathrm{Exp}(\lambda)$ then $X \sim \mathrm{Ga}(\alpha, \lambda)$
- $Ga(1, \lambda) = Exp(\lambda)$

7.5 Beta Distribution

Notation	$X \sim \mathrm{Beta}(\alpha, \beta)$
Experiment	-
Support	$x \in [0,1]$
$f_X(x)$	$\begin{cases} \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} & x \in [0,1] \\ 0 & \text{else} \end{cases}$ $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
$F_X(x)$	$\int_0^x f_X(t)dt$
$\mathbb{E}[X]$	$\dfrac{\alpha}{\alpha+\beta}$
$\mathrm{Var}[X]$	$\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

7.6 χ^2 Distribution

Notation	$X \sim \chi^2(k)$
Experiment	-
Support	$x \in [0, \infty)$ or $x \in (0, \infty)$ if $k = 1$
$f_X(x)$	$\begin{cases} \dfrac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} & x \ge 0 \\ 0 & x < 0 \end{cases}$
$F_X(x)$	$\int_{-\infty}^x f_X(t)dt$
$\mathbb{E}[X]$	k
$\mathrm{Var}[X]$	$2k$

Properties

- Let $X_1, \dots, X_n$ i.i.d. $X_i \sim \mathcal{N}(0,1)$ then $Y = \sum_{i=1}^n X_i^2 \sim \chi^2(n)$
- $X \sim \chi^2(n) \Leftrightarrow X \sim \mathrm{Ga}(\alpha = \frac{n}{2}, \lambda = \frac{1}{2})$

7.7 t-Distribution

Notation	$X \sim \mathrm{t}(n)$
Experiment	-
Support	$x \in \mathbb{R}$
$f_X(x)$	$\dfrac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \dfrac{x^2}{n}\right)^{-\frac{n+1}{2}}$
$F_X(x)$	$t_{n,x}$ *(use t-table)*
$\mathbb{E}[X]$	0
$\mathrm{Var}[X]$	$\dfrac{n}{n-2}$

Properties

- $X \sim \mathrm{t}(n = 1) \Rightarrow X \sim \mathrm{Cauchy}$
- $X \sim \mathrm{t}(n \to \infty) \Rightarrow X \sim \mathcal{N}(0,1)$
- If $n > 30$ we can usually approximate the t-distribution with a normal distribution.

7.8 Cauchy Distribution

Notation	$X \sim \mathrm{Cauchy}(t, s)$
Experiment	-
Support	$x \in \mathbb{R}$
$f_X(x)$	$\dfrac{1}{\pi s \left(1 + \left(\frac{x-t}{s}\right)^2\right)}$
$F_X(x)$	$\dfrac{1}{2} + \dfrac{1}{\pi} \arctan\left(\dfrac{x-t}{s}\right)$
$\mathbb{E}[X]$	*undefined*
$\mathrm{Var}[X]$	*undefined*